Locally incremental visual cluster analysis using Markov random field

Zhen Zhou, Li Zhong, Liang Wang*

Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, PR China

A R T I C L E   I N F O

Article history:
Received 22 October 2013
Received in revised form 7 January 2014
Accepted 15 January 2014
Communicated by L. Shao
Available online 17 February 2014

Keywords:
Visual cluster analysis
Markov random field
Visual assessment tendency

A B S T R A C T

Clustering methods are widely deployed in the fields of data mining and pattern recognition. Many of them require the number of clusters as the input, which may not be practical when it is totally unknown. Several existing visual methods for cluster tendency assessment can be used to estimate the number of clusters by displaying the pairwise dissimilarity matrix into an intensity image where objects are reordered to reveal the hidden data structure as dark blocks along the diagonal. A major limitation of the existing methods is that they are not capable to highlight cluster structure with complex clusters. To address this problem, this paper proposes an effective approach by using Markov Random Fields, which updates each object with its local information dynamically and maximizes the global probability measure. The proposed method can be used to determine the cluster tendency and partition data simultaneously. Experimental results on synthetic and real-world datasets demonstrate the effectiveness of the proposed method.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Clustering is a fundamental technique in pattern recognition and data mining [1,2]. The principle is to maximize inter-cluster difference and minimize intra-cluster difference. Given a set of data points, there are two major clustering approaches. One is hard clustering, for example, K-means [3]. The other is fuzzy clustering, for example, fuzzy C-means [4]. For those methods requiring the number of clusters as the input, the first step is to estimate the cluster tendency.

Visual clustering methods are widely used to assess the cluster tendency, which generally require the pairwise dissimilarity matrix as the input. A popular method is Visual Assessment of cluster Tendency (VAT) [5]. The method produces an intensity image, called Ordered Dissimilarity Image (RDI) or Reordered Dissimilarity Image (RDI) [6], which is able to reveal cluster tendency after suitably reordering the dissimilarity matrix [7].

Several algorithms extend the VAT method. bigVAT [8] and sVAT [9] offer different ways to approximate RDIs for large scale datasets. CCE [10] and DBE [11] use different schemes to automatically estimate the number of clusters in RDIs. It has been found that VAT is effective when data contains compact structures. However, many real datasets involve highly irregular structures.

In this paper, we propose a VAT based approach by using Markov Random Field (MRF) [12] to handle with complex data structures. Meanwhile, the membership matrix is computed simultaneously, which can be directly used for fuzzy clustering. Experimental results on synthesis and real datasets have demonstrated the effectiveness of the proposed method. The remainder of this paper is organized as follows. Section 2 briefly reviews the VAT algorithm and its extensions. The proposed method is introduced in Section 3. The experimental results are displayed and analyzed in Section 4, prior to the conclusion in Section 5.

2. VAT and its extensions

Suppose $X = \{x_i\}_{n}$ denotes the set of $n$ data points and $D = \{d_{ij}\}_{n \times n}$ is the dissimilarity matrix, of which each element denotes the difference between two data points, satisfying $d_{ij} = d_{ji}$ and $d_{ii} = 0$. The goal of the VAT method is to find a permutation rule to reorder the rows of the dissimilarity matrix. The reordered dissimilarity matrix, $\hat{D}$, displays as an intensity image. Such $\hat{D}$ is expected to have a block diagonal form. If $x_i$ is a member of a cluster, the corresponding $d_{ij}$ should be in a sub-matrix with low dissimilarity values, which appears as one of the dark blocks along the diagonal of the VAT image. Each dark block can be regard as a potential group. The VAT algorithm can be found in Table 1 taken from [13].
An example of VAT is shown in Fig. 1, in which (left) the scatter plots of data consist of n=4000 points in $\mathbb{R}^2$. These data points are generated from a mixture of bi-variate Gaussian distributions. The dissimilarity metric is computed by the Euclidean distance between every two points. The 4 visually apparent clusters in (left) are displayed by 4 distinct dark blocks along the diagonal in (right), which is the VAT image of the data. Therefore, reordering is necessary to reveal the underlying cluster structure of data, in contrast to the image (middle) of the input pairwise dissimilarity matrix in an original order.

The VAT algorithm performs well on datasets with compact structures. However, such assumption is often hard to satisfy in the real-world task. Fig. 2 is a toy example of this case. To address this problem, several extensions have been proposed, e.g., improved VAT (iVAT) [6] and Spectral VAT (SpecVAT) [13]. The iVAT algorithm is based on the path-transformation (Eq. (1)). Here, $d_{ij}$ represents the weight of the edge between $x_i$ and $x_j$, $P_{ij}$ is the set of all possible paths from $x_i$ to $x_j$, $|p|$ is the number of vertices along path $p$ and $|h|$ is the index of the $h$th vertex along path $p$. For each path $p \in P_{ij}$, the effective dissimilarity between $x_i$ and $x_j$ along $p$ is the maximum of all weights of this path. Reordering the new matrix $D' = [d_{ij}]_{i=1}^{n}$ obtains the iVAT image (Fig. 2 right).

$$d_{ij} = \min_{p \in P_{ij}} \{ \max_{1 \leq h < |p|} d_{ij}[h+1] \}.$$  

The SpecVAT algorithm is based on spectral transformation [13]. The key of SpecVAT is graph embedding, which first calculates a weighted affinity matrix and uses the Laplacian Eigenmap to transform the pairwise dissimilarity matrix, and then applies the VAT algorithm on this transformed matrix.

SpecVAT, which uses graph-embedding techniques, has also two major weaknesses: one is that the qualities of reordered images by SpecVAT largely depend on the selection of the number of eigenvectors; the other is that SpecVAT fails on sparse and uneven datasets, due to its dimensionality reduction method. Besides, these algorithms can be generally used for cluster tendency assessment, not capable for data partition and labeling. Based on those intuitions, we propose a novel method for visual cluster analysis with the Markov Random Field model [12].

The first step of our model is to construct a graph based on the input dataset. In our case, it is a full connected graph, where each data point is regarded as a node and the edge between two nodes is their dissimilarity computed by the Euclidean distance, thus leading to an undirected graphical model. Meanwhile, it is natural to think that the neighbor of a node will provide extra information. If we change the range of the neighbor from several nearest nodes to the entire graph, it obviously provides the spatial information. This is the way to compute the global probability. With more exploration, we find that the graph holds the local Markov property [14]. Second, we construct a scatter graph representation for the dissimilarity matrix. The next step is to build the local update system by the Markov property. Meanwhile, a criterion is set to stop the iteration. When we find the optimal $k$ nearest neighborhood system (later will explain), we will complete data partition.

Here is the details of our method. At first, there are $n$ groups, written as $\Omega = \{\omega_i\}_{i=1}^{n}$ where $\omega_i$ is the label of $x_i$. The membership matrix $B = [b_{ij}]_{i=1}^{n}$ is set to identity matrix, where $b_{ij} = p(\omega_i = j)$.

**Definition 1** is the neighborhood system, which will be used to compute the global probability in **Definition 2**.

**Definition 1.** $X = \{x_i\}_{i=1}^{n}$ is the set of nodes and $S = (S_t \subseteq X | x_i \in X)$ is a family of subsets of $X$. Note that $S_t$ is the neighborhood of $x_t$ and $x_t \notin S_t$. $S$ is called a neighborhood system, if $\forall t, t \in X$, $t \in S_t \iff r \in S_t$. That means the neighborhood relation is anti-reflexive and symmetric, but not transitive [15].
maximal clique, which is displayed as

\[ D = \{d_{ij}\}; \text{ an } n \times n \text{ scaled matrix of pairwise dissimilarities, with } 1 \geq d_{ii} \geq 0; d_{ij} = d_{ji}; d_{ij} = 0, \text{ for } 1 \leq i, j \leq n \]

(1) Initialization:
Set the membership matrix: \( B = B' \) if we have prior probability information \( B' \); or \( B = I \), which means every node tends to belong to itself and the group number is \( n \) at this moment. Set \( k = 2 \) and the maximum number of iterations equal to \( \log(n) \).

(2) Iteration:
1. \( \forall x_k \in X \), construct the neighborhood system \( I_k^x \).
2. Calculate the global probability \( P \) by Eq. (3), if it reaches the maximum number of iterations or the current \( P \) is no more than the previous one, stop; else \( k \leftarrow 2 \times k \), continue.

(3) Partition the dataset:
Calculate each node’s \( b_k \) by Eq. (4) using the \( k \) obtained from the previous step, which is the degree of the node belonging to the group. This separates the dataset into several groups. Pick up \( j = \arg \max \limits_k b_k \) for each node \( i \) as its group belonging to.

(4) Graph embedding and reordering:
Choose \( I_k^x \) as the new unit and reorder these units according to their unit belongings. Transform the dissimilarity matrix \( D_{n \times n} \) into \( D' \), of which each element is the dissimilarity between two \( k \) nearest neighborhoods.

(5) Implement path-based transformation to \( D' \) by Eq. (1).

(6) Re-transform \( D' \) into \( D''_{n \times n} \) and apply the VAT algorithm to \( D''_{n \times n} \).

Output: Reordered \( D'' \) and its corresponding scaled gray-scale image \( I(D'') \). Membership matrix \( B \) and its separation result.

**Definition 2.** An undirected graphical model is said to be a Markov Random Field if two nodes are conditionally independent whenever they are separated by other nodes. In our case, this means that the following conditional probability holds

\[ P(\omega_0 | \Omega_{x_0}) = P(\omega_0 | \Omega_{x_0} \Gamma) \quad (2) \]

where \( \Omega_{x_0} \) is the labels of all nodes except \( x_0 \) and \( \Omega_{x_0} \) denotes the labels of the neighborhood of \( x_0 \), i.e., all the nodes that are connected to \( x_0 \). Hammersley Clifford Theorem [16] guarantees to compute \( P(\Omega) \) on maximal cliques that cover all the nodes and edges of the graph. The maximal clique is described in **Definition 3**. In order to compute the probability \( P(\Omega) \), we just need to conduct on every maximal clique.

**Definition 3.** For an undirected graph, a clique is a set of nodes that are all neighbors of each other. A maximal clique is a clique if it contains the largest possible number of nodes without losing the clique property.

However, in our method, the graph is a full connected one, which means that the maximal clique of each node is the whole graph, which makes the computational cost unbearable. Therefore, a so-called \( k \) nearest neighborhood is proposed to approximate the maximal clique, which is displayed as \( \Gamma^k = \{ \Gamma^k_x \subseteq \Gamma | x_k \in X \} \). \( \Gamma^k_x \) is the set of \( k \) nearest nodes of \( x_k \). So the formula in Hammersley Clifford Theorem can be transformed as

\[ P(\Omega) = \frac{1}{Z} \prod_{x_k} P(\Omega_{\Gamma^k_x}), \quad Z = \sum_{\forall \Omega_{\Gamma^k_x}} \prod_{x_k} P(\Omega_{\Gamma^k_x}), \quad (3) \]

where \( \Omega_{\Gamma^k_x} \) is the set of labels of \( \Gamma^k_x \), and \( Z \) is a normalized constant which makes it as a probability distribution. Note that \( P \) is monotonically nondecreasing. As we use the \( k \) nearest neighborhood system to approximate the maximal clique, the original Hammersley Clifford Theorem cannot be hold strictly. But it is easy to get that this approximation, compared with the formal one, has an infinitesimal value of \( k \) and \( n \). And this difference becomes smaller and smaller with the increase of \( k \). In order to automatically determine \( k \), our method is to maximize this probability, i.e., maximize on the \( k \) nearest neighborhood system.

When we get the optimal \( k \), the next step is to partition the dataset by using the \( k \) nearest neighborhood system. Consider the meaning of \( b_{ij} = p(\omega_i = j) \), it is natural to think that if the neighbors of \( x_i \) tend to be clustered to group \( j \), \( x_i \) should have the higher probability belonging to that group than others. In addition, the distance between nodes should be considered to emphasize the importance of spatial location. This idea is formalized as

\[ b_{ij} = \frac{1}{Z} \sum_{x_k \in I^k_x} \frac{b_{mj}}{\|x_i - x_m\|^2}, \quad Z = \sum_{x_k \in I^k_x} \sum_{x_m \in I^k_x} \frac{b_{mj}}{\|x_i - x_m\|^2} \quad (4) \]

where \( \|x_i - x_j\|^2 \) denotes the Euclidean distance between \( x_i \) and \( x_j \) and \( Z \) is a normalized constant which makes \( b_{ij} \) as a probability distribution. Consequently, the probability of \( x_i \) belonging to group \( j \), i.e., \( b_{ij} \), is influenced by its \( k \) nearest neighborhood and the pairwise distance within the neighborhood. That means if a node is close to \( x_i \) and has high probability to be clustered to group \( j \), it will have more importance for supporting \( x_i \) belonging to group \( j \). Theoretically, optimal \( k \) makes it possible to combine both kind of global and local information. By now, the motivation and two main.

**Table 2**
The proposed MrIVAT method.

| Input: D = {d_{ij}}; an n × n scaled matrix of pairwise dissimilarities, with 1 ≥ d_{ii} ≥ 0; d_{ij} = d_{ji}; d_{ij} = 0, for 1 ≤ i, j ≤ n |
| (1) Initialization: Set the membership matrix: B = B' if we have prior probability information B'; or B = I, which means every node tends to belong to itself and the group number is n at this moment. Set k = 2 and the maximum number of iterations equal to log(n). |
| (2) Iteration: |
| 1. \( \forall x_k \in X \), construct the neighborhood system \( I_k^x \). |
| 2. Calculate the global probability P by Eq. (3), if it reaches the maximum number of iterations or the current P is no more than the previous one, stop; else \( k \leftarrow 2 \times k \), continue. |
| (3) Partition the dataset: |
| Calculate each node’s \( b_k \) by Eq. (4) using the k obtained from the previous step, which is the degree of the node belonging to the group. This separates the dataset into several groups. Pick up \( j = \arg \max \limits_k b_k \) for each node \( i \) as its group belonging to. |
| (4) Graph embedding and reordering: |
| Choose \( I_k^x \) as the new unit and reorder these units according to their unit belongings. Transform the dissimilarity matrix \( D_{n \times n} \) into \( D' \), of which each element is the dissimilarity between two \( k \) nearest neighborhoods. |
| (5) Implement path-based transformation to \( D' \) by Eq. (1). |
| (6) Re-transform \( D' \) into \( D''_{n \times n} \) and apply the VAT algorithm to \( D''_{n \times n} \). |
| Output: Reordered \( D'' \) and its corresponding scaled gray-scale image \( I(D'') \). Membership matrix \( B \) and its separation result. |

**Fig. 2.** The IVAT algorithm can still estimate the three clusters (right) of the three-line dataset (left), while VAT fails (middle).
procedures of our method have been introduced in detail, which is again summarized in Table 2.

Step (4), the dissimilarity between two \(k\) nearest neighborhoods is computed by the Euclidean distance of their centers. Meanwhile, record the minimal distance within each \(k\) nearest neighborhood, which will be used to reconstruct \(\tilde{D}\) from \(D^0\) in Step (6). Step (5), i.e., path-based transformation, aims to improve the final output image quality. Note that in our model, the center of each cluster is an existing data point, not like \(K\)-means as a virtual one. Step (6), in order to reconstruct \(\tilde{D^0}\) from \(D^0\), every row and column in \(D^0\) will be copied \(k\) times with pending values, which are the minimal distance mentioned in Step (4) if the corresponding two elements are in the same neighborhood, or the distance between the centers of corresponding \(k\) nearest neighborhoods. This extended new matrix is \(\tilde{D^0}\).

4. Experiments

In order to evaluate the performance of the proposed method, we carry out a series of experiments. We compare it with the VAT, iVAT and SpecVAT. The datasets used in this section are 9 synthesis sets and 6 real-world datasets. Unless otherwise mentioned, the dissimilarity matrix used as the input in these experiments is derived from the original attribute space computed by the Euclidean measure.

4.1. Datasets

**Synthetic data**: We use totally 9 synthesis datasets and 5 of them (S1–S5) are from [17], and the remaining datasets (S6–S9) are generated by Karypis Lab for evaluating the CHAMELEON clustering algorithm [18]. Each dataset is composed of hundreds or thousands 2D objects; accordingly, they can be easily visualized using 2D data plots (see Fig. 3 Top and Fig. 4 Left).

**Real data**: Six real-world datasets were also considered to evaluate our algorithm, four of which are from UCI Machine Learning Repository [19]. These datasets are summarized as follows: (a) R-1: Breast-cancer database includes 699 instances, each of which has 9 attributes and belongs to one of 2 classes. (b) R-2: This dataset [20] contains single-light-source Face images of 3 different individuals, each seen under 585 viewing conditions. Each original image is down-sampled to \(30 \times 40\) pixels, thus providing in total 1755 images. (c) R-3: Genetic dataset is originally from the work in [21], which is a \(194 \times 194\) matrix consisting of pairwise dissimilarities from a set of 194 human gene products that were clustered into three protein families. (d) R-4: Iris dataset contains 3 physical classes, 50 instances each, where each class

![Fig. 3](image-url). The first row is the scatter plots of S1–S5. The following 4 rows respectively give the results of the VAT, iVAT, SpecVAT, and MrfVAT.
Fig. 4. The first column is the scatter plots of S6–S9. The following 4 columns respectively give the results of the VAT, iVAT, SpecVAT, and MrfVAT. The real number of clusters of the 4 datasets are 6, 6, 9 and 8, respectively.

Fig. 5. This figure shows the results of the VAT, iVAT, SpecVAT, MrfVAT on R1–R6. The first row represents the results of the VAT, second for the iVAT, third for the SpecVAT, and the last for the MrfVAT. The real numbers of the clusters are 2, 3, 3, 2 and 3, respectively.
refers to a type of iris plant. (e) R-5: Voting dataset consists of 435 US House of Representatives members’ votes on 16 key votes (267 democrats and 168 republicans). (f) R-6: Wine dataset contains the results of a chemical analysis of wines grown in the same region, but derived from 3 different cultivars. The analysis determines the quantities of 13 constituents found in each of three types of wines.

4.2. Visual effects

In Fig. 3, the scatter plots of the first 5 synthesis datasets are displayed in the first row. The following 4 rows respectively give the results of the VAT, iVAT, SpecVAT, and MrfVAT. Generally, the VAT performs well on S3 and S5 datasets while SpecVAT on S2, S3, S4 and S5. The iVAT and MrfVAT are very close in all 5 datasets with clear data structures.

Fig. 4 shows the results of more complicated datasets, S6-S9, all of which have cluttered background (or noise). In the first column, there are 6 characters in the first scatter plot, 6 in the second, 9 in the third and 8 in the last. The following 4 columns respectively give the results of the VAT, iVAT, SpecVAT, and MrfVAT. These results show the ability of the MrfVAT when dealing with datasets of such complicated structures.

Fig. 5 displays the results of real-world datasets. The first row represents the results of the VAT, second for the iVAT, third for the SpecVAT, and the last for the MrfVAT. The actual number of clusters of each dataset is 2, 3, 3, 2 and 3, respectively. The MrfVAT and SpecVAT are able to reveal correct numbers on four datasets. We can see that MrfVAT can correctly estimate the cluster number on most datasets, even though the estimation may not be the real number of clusters, it would be very close and its images are of high quality, thus providing useful information of cluster tendency.

### 4.3. Clustering

Naturally, we quantify the performance of the proposed method by using the accuracy of clustering. This metric is widely used for evaluating clustering performance [22–24]. The K-means [3] and Ward’s hierarchy [25] methods are also displayed to show the results of classical clustering algorithms. The performances have been summarized in Table 3, from which it can be concluded that the proposed method is comparable to the SpecVAT and outperforms the other methods on the testing datasets. The advantage of the proposed method is that it can deal with datasets with complicated structures, which are displayed in Fig. 4. The important thing is that the visual effect and clustering accuracy of the proposed method may be inconsistent. This is because the partition procedure of the proposed method is independent of the subsequent operator, i.e., reordering the dissimilarity matrix, although they share the same number of clusters.

### 4.4. Image segmentation

Next, we apply the proposed method to the problem of high-resolution image segmentation, which is always large scale and infeasible to directly handle with the available computing resource. A sampling-based scheme is generally necessary, which can be found in [26]. To perform visual cluster analysis on such large scale datasets, we apply the proposed method to representative samples, and then obtain the clustering results of the remaining samples, just like [27,28,26]. Fig. 6 shows the segmentation results of the used images, which are also used in [26]. It is obvious that the proposed method partitions the images into meaningful components, which demonstrates the effectiveness of our method.

### 4.5. Complexity and memory consumption

Last but not the least, the complexity of the VAT is $O(n^2)$ while the ones of iVAT and SpecVAT are of the same magnitude, $O(n^2)$ [29]. In the proposed method, it has a complexity of $O(n \log(n))$ before Step (5) and finally $O(n^3)$.

### 5. Discussion and conclusion

The proposed MrfVAT method is based on the Markov property which is computed by the $k$ nearest neighborhood system. It takes spatial information into account, thus leading to more reasonable partition. In the proposed MrfVAT, data partition is running simultaneously. The membership matrix $B$ makes fuzzy clustering
possible. The experimental results are sufficient to demonstrate the effectiveness of the proposed method.

In future, our work will focus on extending the VAT-based methods to deal with directed graph. Currently, most VAT-based methods can be only applied on undirected graph. However, the information provided by links between vertices, especially the directional ones, can be very useful to predict the behavior of the vertices. Such work is designed for data coming from social networks.

Acknowledgements

This work is jointly supported by National Natural Science Foundation of China (61175003, 61203252), Hundred Talents Program of CAS and National Basic Research Program of China (2012CB316300).

References